Criteria for NTP Reproductive Toxicology Studies

Report from the
Reproductive and Developmental Criteria Working Group (RDCWG)
of the
NTP Board of Scientific Counselors

Submitted by:
Edward Carney, Ph.D., chair
Kim Boekelheide, M.D., Ph.D., rapporteur

At the National Toxicology Program (NTP) Board of Scientific Counselors (BSC) meeting on June 12 2008, Dr. Paul Foster, NTP, provided an outline of the criteria used by the NTP to describe the results of the carcinogenesis bioassay and briefly discussed the NTP's plan to develop similar criteria for reproductive and developmental toxicology studies. The NTP proposed to form working groups to formulate these criteria. Thus, the purpose of the Reproductive and Developmental Criteria Working Group (RDCWG) was to investigate the utility of having specific criteria for describing the results from individual NTP reproductive and developmental toxicology reports to indicate the strength of the evidence for their conclusions. The RDCWG was composed of ten scientists representing academia, industry, and government. Dr. Edward Carney, The Dow Chemical Company, a member of the NTP BSC, chaired the RTCWG. Drs. Barry Delclos, National Center for Toxicological Research/NTP, Dr. Mark Cesta, National Institute of Environmental Health Sciences/NTP and Dr. Paul Foster, Acting Branch Chief, Toxicology Branch served as advisors to the RDCWG. Drs. Kim Boekelheide, Brown University, and Barbara Shane, NTP Executive Secretary, served as rapporteurs. Also attending the meeting from the NTP was Dr. Mary Wolfe, NTP Federal official. The full RDCWG roster is attached [Appendix A]. The RTDWG met September 10 and 11, 2008 at the Hilton Garden Inn Durham/Southpoint Hotel, 7007 Fayetteville Road, Durham, NC.

The NTP developed draft criteria for describing results of NTP reproductive and developmental studies that were modeled after the NTP criteria used to evaluate carcinogenicity studies. Dr. Foster was the lead scientist for this effort. Prior to the RDCWG meeting, the draft criteria were evaluated internally. The RDCWG was tasked to first evaluate the draft criteria for reproductive toxicology studies and then draft criteria for developmental toxicology studies. This report addresses the revision and discussion by the RDCWG regarding the draft criteria for NTP reproductive toxicology studies. A separate report was prepared to discuss the RDCWG's evaluation of the draft criteria for NTP developmental toxicology studies.

Dr. Foster opened the meeting by providing the background for the development of the criteria by NTP. He presented information regarding NTP's reproductive toxicology testing strategies and a discussion of the reproductive toxicology criteria. Materials provided to the RDCWG included: the draft criteria [Attachment B], a set of case studies for testing the utility and applicability of the draft criteria for reaching conclusions on NTP developmental toxicology studies [Attachment C], a list of issues for discussion by the RDCWG [Attachment D], and the carcinogenicity criteria [Attachment E]. The RDCWG was given the following charge:

> *Evaluate the suitability and utility of the proposed criteria for describing the results from individual NTP* reproductive toxicology *studies to indicate the strength of the evidence for their conclusions.*

The RDCWG completed the case study exercise, deliberated on the proposed criteria, and produced the following revised criteria based on those discussions. In revising the draft criteria, the RDCWG deliberated a number of issues that are discussed below (see

“RDCWG Discussion”). Their deliberations resulted in the following revised draft criteria:

EXPLANATION OF LEVELS OF EVIDENCE FOR REPRODUCTIVE TOXICITY

The NTP describes the results of individual studies of chemical agents, and notes the strength of the evidence for conclusions regarding each study. Generally, each study is confined to a single laboratory animal species, although in some instances, multiple species may be investigated under the purview of a single study report. Negative results, in which the study animals do not exhibit evidence of reproductive toxicity, do not necessarily imply that a chemical is not a reproductive toxicant, but only that the chemical is not a reproductive toxicant under these specific conditions. Positive results demonstrating that a chemical causes reproductive toxicity in laboratory animals under the conditions of the study are assumed to be relevant to humans, unless data are available which demonstrate otherwise. In addition, such positive effects should be assumed to be primary effects, unless there is clear evidence that they are secondary consequences of excessive toxicity to non-reproductive organ systems. Given that developmental events are intertwined in the reproductive process, effects on developmental toxicity may be detected in reproductive studies. Evaluation of such developmental effects should be based on the NTP Criteria for Levels of Evidence for Developmental Toxicity.

It is critical to recognize that the “levels of evidence” statements described herein only describe reproductive **hazard**. The actual determination of **risk** to humans requires exposure data that are not considered in these summary statements. This fact is particularly important to keep in mind when communicating study results to the general public.

Five categories of evidence of reproductive toxicity are used in the NTP Technical Report series to summarize the strength of the evidence observed in each experiment: two categories for positive results **(clear evidence and some evidence);** one category for uncertain findings **(equivocal evidence);** one category for no observable effects **(no evidence);** and one category for experiments that cannot be evaluated because of major design or performance flaws **(inadequate study).** In addition, the study’s lowest observed adverse effect level is reported for positive results, and the highest dose level tested is reported for the **no evidence** category. Application of these criteria requires professional judgment by individuals with ample experience with and understanding of the animal models and study designs employed. For each study, conclusion statements are made using one of the following five categories to describe the findings; if warranted, these conclusion statements should be made separately for males and females. These categories refer to the weight of evidence of the experimental results and not to potency or mechanism.

- **Clear evidence** of reproductive toxicity is demonstrated by a dose-related[1] effect on fertility or fecundity, or by changes in multiple interrelated reproductive parameters of sufficient magnitude that by weight of evidence implies a compromise in reproductive function. A statement to the effect of "This study has a lowest observed adverse effect level of XXXX mg/kg/d or other appropriate units (e.g. ppm in diet, mg/L in drinking water) for reproductive toxicity" should accompany the evidence statement.

- **Some evidence** of reproductive toxicity is demonstrated by deficits in reproductive parameters, the net impact of which is judged by weight of evidence to have potential to compromise reproductive function. Relative to clear evidence of reproductive toxicity, such effects would be characterized by greater uncertainties or weaker relationships with regard to dose, severity, magnitude, incidence, persistence and/or decreased concordance among affected endpoints. A statement to the effect of "This study has a lowest observed adverse effect level of XXXX mg/kg/d or other appropriate units (e.g. ppm in diet, mg/L in drinking water) for reproductive toxicity" should accompany the evidence statement, except in those instances in which the "some" classification has been based on uncertainties about the dose relationship that precludes confident determination of the LOAEL.

- **Equivocal evidence** of reproductive toxicity is demonstrated by marginal or discordant deficits in reproductive parameters that may or may not be related to the test article.

.

- **No evidence** of reproductive toxicity is demonstrated by data from a well conducted, adequate study that are interpreted as showing no biologically relevant evidence of chemically-related deficits in reproductive parameters. A statement to the effect of "This study had no observable adverse reproductive toxicity at the highest dose tested (XXXX mg/kg/d or other appropriate units (e.g. ppm in diet, mg/L in drinking water)" should accompany the evidence statement.

- **Inadequate study** of reproductive toxicity is demonstrated by a study that, because of major design or performance flaws, cannot be used to determine the presence of reproductive toxicity.

Other Key Points for Consideration

When a conclusion statement for a particular experiment is selected, consideration must be given to key factors that would extend the boundary of an individual category of evidence. Such consideration should allow for incorporation of scientific experience and

[1]The term "dose-related" describes any dose relationship, recognizing that the treatment-related responses for some endpoints may be non-monotonic due to saturation of exposure or effect, overlapping dose-response behaviors, change in manifestation of the effect at different dose levels, or other phenomena.

current understanding of reproductive toxicity studies in laboratory animals, particularly with respect to interrelationships between endpoints, impact of the change on reproductive function, relative sensitivity of end points, normal background incidence, and specificity of the effect. For those evaluations that may be on the borderline between two adjacent levels, some factors to consider in selecting the level of evidence of reproductive toxicity are given below:

- Increases in severity and/or prevalence (more individuals and/or more litters) as a function of dose generally strengthen the level of evidence, keeping in mind that the specific manifestation may change with increasing dose. For example, histological changes at a lower dose level may reflect reductions in fertility at higher dose levels.

- In general, the more animals affected, the stronger the evidence; however, effects on a small number of animals across multiple related endpoints should not be discounted, even in the absence of statistical significance for the individual endpoint(s). In addition, malformations with low incidence should be interpreted in the context of historical controls and may be biologically important.

- Consistency of effects across generations strengthens the level of evidence. However, special care should be taken for decrements in reproductive parameters noted in the F1 generation that were not seen in the F0 generation, which may suggest developmental as well as reproductive toxicity. Alternatively, if effects are observed in the F1 generation but not in the F2 generation (or the effects occur at a lesser frequency in the F2 generation), this may be due to the nature of the effect resulting in selection (i.e., if the effect is incompatible with successful reproduction, then the affected individuals will not produce offspring).

- Transient changes (e.g., pup weight decrements) by themselves are weaker indicators of effect than persistent changes.

- Single endpoint changes by themselves are weaker indicators of effect than concordant effects on multiple interrelated endpoints.

- Insights from supportive studies (e.g., toxicokinetics, ADME, computational models, structure-activity relationships) and reproductive findings from other *in vivo* animal studies (NTP or otherwise) should be drawn upon when interpreting the biological plausibility of a change.

- New technical approaches and highly sensitive techniques need to have been appropriately characterized to build confidence in their utility, and their usefulness as indicators of effect is increased if they have been anchored to changes in traditional endpoints.

- Clear changes in multiple reproductive tract endpoints without functional changes are sufficient to reach a conclusion of clear evidence of reproductive toxicity

Ancillary recommendations

For **Some/Equivocal Evidence** calls or discordant effects, it may be important to convey whether additional studies are needed to clarify the effects. If additional studies are recommended, then the specific area of concern and recommended approach should be articulated.

RDCWG Discussion

The RDCWG deliberated a number of issues in determining what revisions were needed to the draft criteria and the factors that should be considered when determining the appropriate level of evidence for describing a study's results. This was a rich discussion that resulted in agreement by the Working Group on the levels of evidence, and on the bullet items listed in the sections above entitled "Other Key Points for Consideration" and "Ancillary Recommendations."

As described earlier, the process by which the RDCWG addressed its charge began with a draft ("straw man") criteria document provided by NTP. The RDCWG agreed with the general approach of structuring the proposed criteria after the carcinogenicity criteria and agreed that the proposed number of categories was appropriate. There was significant discussion about the hazard-based nature of the summary statements, with concerns expressed about the general public's tendency to view hazard as synonymous with risk. As such, the RDCWG asks NTP and other end users of the criteria documents to use adequate caution when using the criteria and summary statements to communicate to the general public. Along similar lines, there was a vibrant discussion on whether or not to include some indication as to the dose level required to elicit adverse reproductive effects, as many felt that this information is fundamental to the characterization of a chemical's potential hazard. The RDCWG recognized the need to communicate this dose level information in a simple, coherent manner, leading to the recommendation that a short statement declaring the LOEL (for "clear evidence" or "some evidence" categories) or NOEL (for "no evidence category") accompany the summary statement.

Beyond these general issues, much of the discussion was intended to refine the specific wording of the criteria and was driven by the case studies. These case studies were provided by both NTP and members of the RDCWG and were purposely designed to reside in the "transition zones" between categories. The case studies initially were reviewed and scored separately by each RDCWG member and the results tallied so the group could view the degree of concordance (or lack thereof). The ensuing discussions revealed the thought process behind each member's score, and proved quite constructive in refining the criteria so that the boundaries between categories were as clear as reasonably possible. However, as the interpretation of complex data always carries with it some degree of judgement, it is recommended that NTP develop some additional examples by which the criteria were applied in order to accumulate some “case history”. These examples,which could be developed over time should be appended to the Criteria, and would supplement and exemplify the specific criteria and considerations adopted in the criteria document. An additional benefit of having these examples appended to the criteria is that as scientific knowledge accumulates over time and the boundaries between these categories drift, these examples could provide tracking of those changes.

Finally, some additional points were discussed which are somewhat ancillary to the criteria, but which nonetheless may be of interest to NTP. One of these relates to primordial follicle count, which is a required end point in the OECD and EPA 2-generation study test guideline. There was a strong sentiment among Working Group members that primordial follicle counts are highly variable and a weak indicator of effect. Consistent with recommendations recently published by the Society of Toxicological Pathologists, primordial follicle counts are not considered to be of value in standard reproductive toxicology studies, such as the 2-generation study. This end point is better suited for more targeted investigative research studies.

Appendix A

NTP Board of Scientific Counselors Reproductive and Developmental Criteria Working Group

Working Group Members

Kim Boekelheide, M.D., Ph.D.
Professor, Division of Biology and Medicine
Brown University
70 Ship Street
(Chestnut Street Loading Dock)
Providence, RI 02903

Tracie Bunton, D.V.M., Ph.D., D.A.C.V.P.
Pharmaceutical Toxicology and Pathology Consulting
EICARTE LLC
c/o 150 Irishtown Road
Fairfield, PA 17320

Edward Carney, Ph.D. (Chair)
Technical Leader, Development Reproductive and General Toxicology
The Dow Chemical Company
Building 1803
Midland, MI 48674

Robert Chapin, Ph.D.
Pfizer
Eastern Point Road, Bldg 274
Groton, CT 06340

George Daston, Ph.D.
Miami Valley laboratories
The Procter and Gamble Company
11810 E. Miami River Rd.
Cincinnati, OH 45253

James M. Donald, Ph.D
Chief, Reproductive Toxicology and Epidemiology Section
Reproductive and Cancer Hazard Assessment Branch
Office of Environmental Health Hazard Assessment
1001 I Street, P.O. Box 4010, MS 12B
Sacramento, CA 95812

L. Earl Gray, Ph.D.
USEPA
NHEERL, Reproductive Toxicology Division
Endocrinology Branch
EB (MD-72)
Research Triangle Park, NC 27711

Barry McIntyre, Ph.D., D.A.B.T.
Reproductive Toxicology
Safety Evaluation Center
Schering-Plough Research Institute
556 Morris Avenue, Bldg. 12
Summit, NJ 07901

Kenneth M. Portier, Ph.D.
Director of Statistics
Statistics and Evaluation Center
Research Department
American Cancer Society
250 Williams Street, Suite 600
Atlanta, GA 30303

Shelley Tyl, Ph.D.
Center for Life Sciences and Toxicology
RTI International
Hermann Laboratory Building, Room 124
3040 Cornwallis Road
Research Triangle Park, NC 27709

Technical Advisors

Mark Cesta, D.V.M., D.A.C.V.P.
Cellular and Molecular Pathology Branch
National Toxicology Program
National Institute of Environmental
Health Sciences
P.O. Box 12233, MD B3-06
Research Triangle Park, NC 27709

Barry Delclos, Ph.D.
Department of Biochemical Toxicology
United States Food and Drug Administration
National Center for Toxicological Research
3900 NCTR Road HFT 110
Jefferson, AR 72079

Paul Foster, Ph.D.
Toxicology Branch
National Toxicology Program
National Institute of Environmental Health Sciences
P.O. Box 12233, MD EC-34
Research Triangle Park, NC 27709

NTP Executive Secretary

Barbara Shane, Ph.D., D.A.B.T.
Office of Liaison, Policy, and Review
National Institute of Environmental Health Sciences
P.O. Box 12233, MD A3-01
Research Triangle Park, NC 27709

NTP Federal Official

Mary S. Wolfe, Ph.D.
Deputy Program Director for Policy
Director, NTP Office of Liaison, Policy, and Review
National Toxicology Program
National Institute of Environmental Health Sciences
P.O. Box 12233, MD EC-31
Research Triangle Park, NC 27709

Levels of Evidence Criteria for Reproductive Toxicity Studies

1. **Clear Evidence of Reproductive Toxicity**

   Demonstrated by the results of a study or studies, in one or more species, that indicate a clear treatment-related effect on fertility and/ or other functional reproductive parameters (e.g., litter size) that is not secondary to overt systemic toxicity or:

   Demonstrated by study results that indicate concordant effects on multiple endpoints that indicate biological plausibility of the response.

2. **Some Evidence of Reproductive Toxicity**

   Demonstrated by a study or studies indicating a chemical related increase in deficits in reproductive parameters in which the strength of the response is less than that required for clear evidence i.e. in the absence of effects on fertility or fecundity.

   For example, there may be statistically significant deficits in the histology of the reproductive organs, but without any clear effects on associated litter parameters (e.g., no changes in percentage of animals pregnant or changes in litter size), or where deficits have been noted in one end point (e.g., a diminution in sperm parameters) without any associated histological changes.

3. **Equivocal Evidence of Reproductive Toxicity**

   Demonstrated by a study or studies that are interpreted as showing marginal deficits in reproductive parameters that may or may not be chemical-related.

4. **No Evidence of Reproductive Toxicity**

   Demonstrated by a well conducted study or studies that are interpreted as showing no biologically relevant evidence of chemically-related deficits in reproductive parameters.

5. **Inadequate Study of Reproductive Toxicity**

   Demonstrated by a study or studies that because of major qualitative or quantitative limitations cannot be interpreted as valid for showing the presence or absence of reproductive toxicity.

Other Key Points for Consideration

- Adequacy of experimental design and conduct.

- Occurrence of common versus rare reproductive deficits.

- Use of historical control data to place concurrent control into perspective and estimate population background incidence of reproductive parameters.

- Concordance of reproductive end points (e.g. Did a decrease in litter size relate to ovarian histology and changes in vaginal cytology?)

- Did the reproductive deficits become more severe with increases in dose? For example, did histological changes at one dose level become decrements in litter size and then reductions in fertility at higher dose levels in any generation?

- Did the reproductive deficits increase in prevalence (more individuals and/or more litters) with dose level in any generation?

- Special care should be taken for decrements in reproductive parameters noted in the F_1 generation (and potentially later generations) that were not seen in the F_0 generation, which may suggest developmental as well as reproductive toxicity.

- The evidence of deficits in reproductive parameters may be supported by other data from *in vivo* animal studies (e.g., 14 and 90-day studies) that may show gross and histopathological effects in the reproductive or associated endocrine organs of a nature and extent that indicates a high likelihood of an adverse effect on reproductive function (e.g., testicular atrophy).

- Effects in one species are sufficient for a conclusion. (We routinely only undertake reproduction studies in one species.)

- Individual conclusion statements should be based on affected sex(es) when determined.

- Professional judgment and understanding of the models employed.

Appendix C

Case Study Exercise for Reproductive Toxicology Studies

Introduction and General Points

The "Levels of Evidence" criteria are loosely based on those used in the NTP toxicology and carcinogenicity reports. All three sets of draft criteria for our non-cancer toxicities that are being evaluated in BSC work groups (reproductive and developmental toxicities and immunotoxicity) have wording about "concordance of end points and biological plausibility," because we are dealing with multiple end points in these toxicities where some are redundant and/or should be linked. Our approach is essentially a weight of evidence type approach – the greater the weight of evidence, the more likely a more severe conclusion will be reached.

Also, note that for reproductive toxicity and immunotoxicity, we are proposing an effect on **integrated function** to meet the criteria for a "clear evidence" designation. Note that the "key considerations/ points" are outlined separately. You may wish to look at these first to aid you in how to consider the data and put the draft criteria into context.

In the case studies, the descriptions are purposely short (to generate some discussion) and provided in a series of bullets. For the purposes of this reproductive toxicity exercise, please assume that these data are from studies that meet (or exceed) the current EPA guidelines for multigenerational studies and are conducted in the rat.

If an effect is noted in the bullets, please assume it is statistically significant and dose-related (unless it is specifically noted otherwise). If effects are not specifically noted, please assume they were not significantly different from controls (and not missing).

As a rule of thumb for these studies, a decrease in terminal body weight greater than 10% between a treated group and controls exceeds the normal amount of systemic toxicity expected at the highest dose level – but beware – if, for example, a test article produced fetal death, then the maternal body weights could be reduced by >10%. This would still be a meaningful toxicological effect, but not necessarily one resulting as a consequence of selecting too high a dose level for the dam.

Case Study # 1

- Parental body weight at top dose level is reduced by 8 % (not significant)
- Small decrease (15% at top dose level) in sperm parameters (motility and count)
- No effect on testis or epididymal morphology or weight
- No effect on litter size or fertile pairings

Case Study # 2

- Small, but statistically significant advance in female puberty (2 days in mid and top dose groups).
- No effect on fertility or litter size
- No effect on ovarian morphology
- ~10% decrease in body weight at highest dose level

Case Study #3

- ~10% decrease in F0 body weight at highest dose level
- Testicular histological findings (retention of spermatids)
- No decrease in testis weight
- Small, non-significant reduction in mean litter size (from 13 to 11)

Case Study #4

- 11% decrease in F0 and F1 body weight at the top dose level
- Decrease in testis and ovary weights (F0)
- No histology changes in the gonads
- No effects on litter parameters or fertility

Case Study #5

- Body weight reduced to 92% of control levels at top dose level
- 3 of 20 pairs not fertile at top dose level (not statistically significant)
- Reduction in mean litter size (13 to 8 at top dose level)
- No effects on male or female gonad histology
- No significant changes in sperm or follicle parameters

Case Study #6A

- No effect on body weight
- Decrease in number of fertile pairings (F0)
- Decrease in mean litter size (F0 and F1)
- Testis histological effects in a few, but not all adults (borderline significance)
- Sperm parameters decreased, but not always significant at all dose levels

Case Study #6B

- Significant (dose-dependent) decrease in pup weight in F1 and F2 generations at PND 4, 7, 14, and 21 at all dose levels, without body weight effects in the adults

Case study # 7

- 12% decrease in parental body weight at highest dose level
- Decrease in number of preantral follicles (F1 and F2)
- No effect on ovarian weight
- No significant decreases in fertility or litter size (non-significant reduction from 13 to 10 at top dose level)

Case study # 8

- No effect on body weight (max. 7% decrease)
- Decrease in epididymal weight
- Small decrease in sperm count (15% at top dose level)
- Decrease in sperm motility (up to 50% in 4 animals at top dose level)
- No fertility or litter effects

Case study # 9

- No effect on adult body weight
- Decrease in pup growth up to PND 12 in top two dose levels, recovered by time of weaning (PND 21)

Case study # 10

- No effect on adult body weight
- Decrease in pup growth after PND 12 in top two dose levels that is still decreased in top dose level by weaning (PND 21)

Case study # 11

- Decrease in body weight at top dose level (15%)
- Significant effect on estrous cyclicity (persistent diestrus) at all dose levels
- No effect on ovarian histology
- No effect on pre-coital interval
- No effect on fertility
- Litter size decreased (from 13 to 9) in the top dose group only

Case study #12

- No effect on body weight
- Decrease in AGD in males (10 % at top dose level)
- 2-day delay in PPS at top dose level
- Small, non-significant increase in uterine malformations in F1 adults: 3/20 at top dose, 0/20 at mid dose, 1/20 at low dose (historical control rate is 1/1700)

Case study #13

- A 9% decrease in parental body weight (in the F0 and F1 generations) at the top dose level Decrease in testis weight (96% control) at the top dose level only (F0 and F1)
- Testicular histology findings (top dose 15% of tubules atrophic; mid dose 5 % tubules atrophic – F0 & F1)
- No effects on fertility or litter parameters
- No effects on sperm parameters

Case study # 14

- Exposure of adult male rats produces a slight, but significant change in body weight at the top dose level
- No decrease in testis weight or epididymal weight at the top dose level
- Testicular histology findings are limited to a dose-dependent increase in retained spermatid heads. Morphometry shows that spermatid head retention is significantly increased in stages IX-XII at the top 2 doses
- Inclusion of a recovery group examined 90 days after stopping exposure shows reversal of the spermatid head retention
- No effects on fertility or litter parameters
- No effects on sperm parameters

Case study #15

- Exposure of adult male rats produces a slight, but significant change in body weight at the top dose level
- No decrease in testis weight or epididymal weight at the top dose level
- Testicular histology findings are limited to a dose-dependent increase in TUNEL positive events. Morphometry shows that increased TUNEL positivity occurs at the top 2 doses
- Inclusion of a recovery group examined 90 days after stopping exposure shows reversal of the TUNEL positivity
- No effects on fertility or litter parameters
- No effects on sperm parameters

Appendix D

Issues for Discussion with NTP BSC Developmental Criteria Working Group

1. Conclusions statements for NTP studies are hazard and not risk-based, to facilitate comparison across chemicals using the same study types. These conclusion statements are voted upon by the NTP Board of Scientific Counselors (BSC) in its advisory role to the NTP Executive Committee, which contains representatives from our sister regulatory agencies that can use this information in quantitative risk assessment decisions.
2. It would be helpful if we could model conclusion criteria for non-cancer studies based on that currently employed for the NTP carcinogenicity studies (attached), to generate some consistency in approach and wording for both the BSC and the public.
3. NTP staff recognizes that for many of the non-cancer toxicity studies, we are dealing with multiple (inter-related) end points very different from cancer studies. Thus, the NTP cancer study approach to levels of evidence in drawing study conclusions will require some "finessing" to achieve the desired level of consistency.
4. NTP staff also recognizes the desirability to use a graded (hazard identification) conclusion scheme, such that a single positive finding does not necessarily result in the highest level of conclusion. We have considered those end points that affect overall function to merit the highest level of conclusion (clear evidence of toxicity). So, there may be a statistically significant, dose-related decrease in some end point (for example, sperm count in a reproduction study), but without a concomitant effect on animal function (e.g., fertility or litter size parameters), it would not merit the clear evidence category.

Appendix E

EXPLANATION OF LEVELS OF EVIDENCE OF CARCINOGENIC ACTIVITY

The National Toxicology Program describes the results of individual experiments on a chemical agent and notes the strength of the evidence for conclusions regarding each study. Negative results, in which the study animals do not have a greater incidence of neoplasia than control animals, do not necessarily mean that a chemical is not a carcinogen, inasmuch as the experiments are conducted under a limited set of conditions. Positive results demonstrate that a chemical is carcinogenic for laboratory animals under the conditions of the study and indicate that exposure to the chemical has the potential for hazard to humans. Other organizations, such as the International Agency for Research on Cancer, assign a strength of evidence for conclusions based on an examination of all available evidence, including animal studies such as those conducted by the NTP, epidemiologic studies, and estimates of exposure. Thus, the actual determination of risk to humans from chemicals found to be carcinogenic in laboratory animals requires a wider analysis that extends beyond the purview of these studies.

Five categories of evidence of carcinogenic activity are used in the Technical Report series to summarize the strength of the evidence observed in each experiment: two categories for positive results **(clear evidence and some evidence);** one category for uncertain findings **(equivocal evidence);** one category for no observable effects **(no evidence);** and one category for experiments that cannot be evaluated because of major flaws **(inadequate study).** These categories of interpretative conclusions were first adopted in June 1983 and then revised in March 1986 for use in the Technical Report series to incorporate more specifically the concept of actual weight of evidence of carcinogenic activity. For each separate experiment (male rats, female rats, male mice, female mice), one of the following five categories is selected to describe the findings. These categories refer to the strength of the experimental evidence and not to potency or mechanism.

- **Clear evidence** of carcinogenic activity is demonstrated by studies that are interpreted as showing a dose-related (i) increase of malignant neoplasms, (ii) increase of a combination of malignant and benign neoplasms, or (iii) marked increase of benign neoplasms if there is an indication from this or other studies of the ability of such tumors to progress to malignancy.
- **Some evidence** of carcinogenic activity is demonstrated by studies that are interpreted as showing a chemical-related increased incidence of neoplasms (malignant, benign, or combined) in which the strength of the response is less than that required for clear evidence.
- **Equivocal evidence** of carcinogenic activity is demonstrated by studies that are interpreted as showing a marginal increase of neoplasms that may be chemical related.
- **No evidence** of carcinogenic activity is demonstrated by studies that are interpreted as showing no chemical-related increases in malignant or benign neoplasms.
- **Inadequate study** of carcinogenic activity is demonstrated by studies that, because of major qualitative or quantitative limitations, cannot be interpreted as valid for showing either the presence or absence of carcinogenic activity.

For studies showing multiple chemical-related neoplastic effects that if considered individually would be assigned to different levels of evidence categories, the following convention has been adopted to convey completely the study results. In a study with clear evidence of carcinogenic activity at some tissue sites, other responses that alone might be deemed some evidence are indicated as “were also related” to chemical exposure. In studies with clear or some evidence of carcinogenic activity, other responses that alone might be termed equivocal evidence are indicated as “may have been” related to chemical exposure.

When a conclusion statement for a particular experiment is selected, consideration must be given to key factors that would extend the actual boundary of an individual category of evidence. Such consideration should allow for incorporation of scientific experience and current understanding of long-term carcinogenesis studies in laboratory animals, especially for those evaluations that may be on the borderline between two adjacent levels.

These considerations should include:

- adequacy of the experimental design and conduct;
- occurrence of common versus uncommon neoplasia;
- progression (or lack thereof) from benign to malignant neoplasia as well as from preneoplastic to neoplastic lesions;
- some benign neoplasms have the capacity to regress but others (of the same morphologic type) progress. At present, it is impossible to identify the difference. Therefore, where progression is known to be a possibility, the most prudent course is to assume that benign neoplasms of those types have the potential to become malignant;
- combining benign and malignant tumor incidence known or thought to represent stages of progression in the same organ or tissue;
- latency in tumor induction;
- multiplicity in site-specific neoplasia;
- metastases;
- supporting information from proliferative lesions (hyperplasia) in the same site of neoplasia or in other experiments (same lesion in another sex or species);
- presence or absence of dose relationships;
- statistical significance of the observed tumor increase;
- concurrent control tumor incidence as well as the historical control rate and variability for a specific neoplasm;
- survival-adjusted analyses and false positive or false negative concerns;
- structure-activity correlations; and
- in some cases, genetic toxicology.